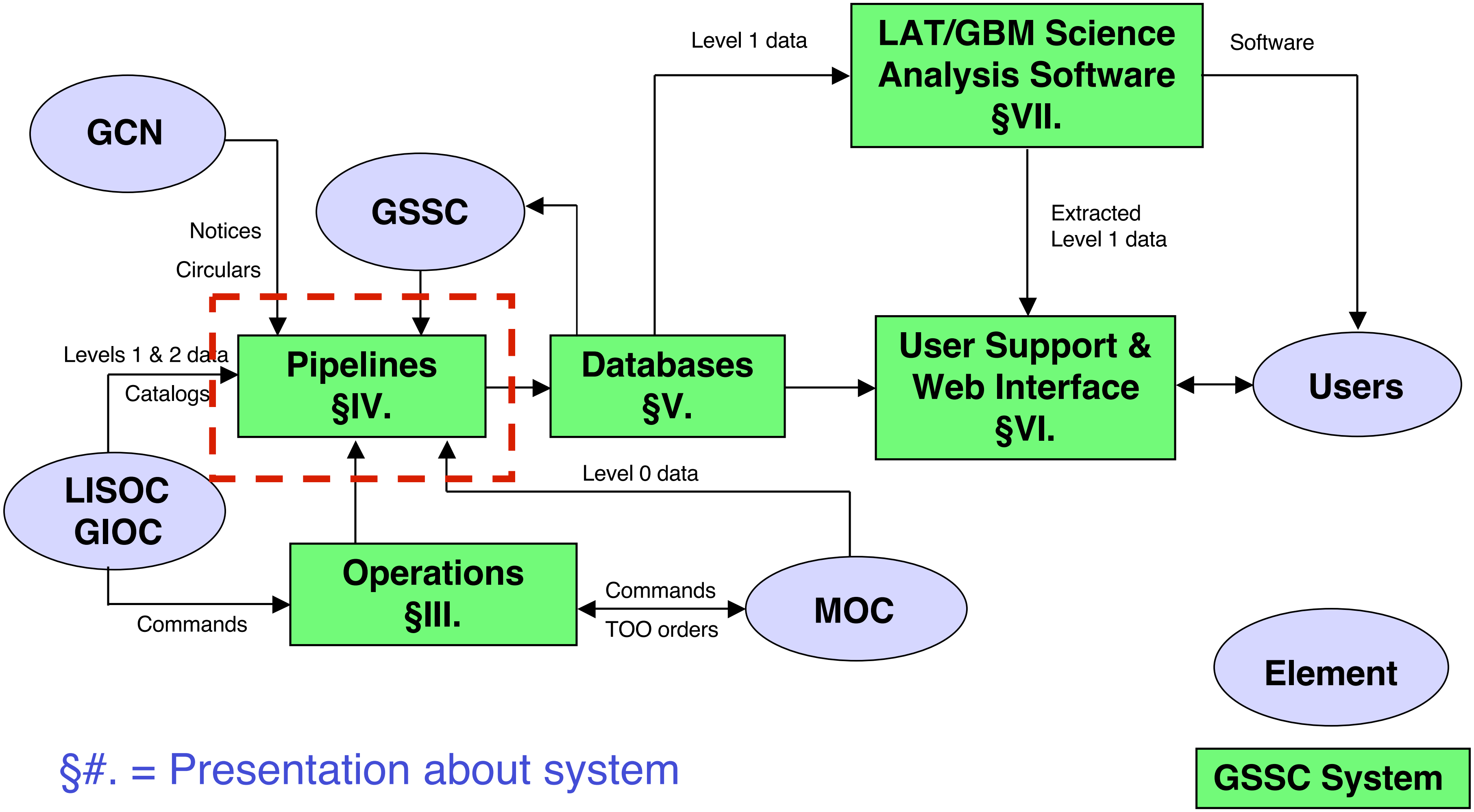
IV – Pipelines

Tom Stephens – GSSC Programmer

Outline

- GSSC Pipelines
- Data Ingest Pipeline
  - Description
  - Requirements
  - Data Transfer System
  - Process Manager
  - Pipeline Manager
  - Specific Pipeline Branches
- Backup Pipelines
- Summary

GSSC Software Systems



§#. = Presentation about system

GSSC Pipelines

- Three data pipelines at the GSSC
  - Data Ingest Pipeline
  - Backup LAT Level 1 Processing Pipeline
  - Backup GBM Level 1 Processing Pipeline
- GSSC responsible for design and implementation of the Data Ingest Pipeline
- Instrument teams responsible for design and development of the Level 1 pipelines – GSSC will host the backup system.

Data Ingest Definition

- The data ingest subsystem is the entry point for all data coming into the GSSC from the rest of the Ground System.

- It consists of a series of programs and scripts that process all arriving data.

Data Ingest Documents

- The requirements and design specifications for the Data Ingest Subsystem come from a variety of documents:
  - GSSC Functional Requirements Document (FRD) 443-RQMT-0002
  - LAT Event Summary Database Requirements Document (LESRD) GSSC-0006
  - Standard Analysis Environment Database Requirements Document (SAEDR) GSSC-0007
  - GSSC Design Document (GDD) GSSC-0003

General Ingest Requirements

General Requirements	
Requirement	Description
FRD 5.4.1.4.1	The GSSC shall interface with the MOC for the exchange of mission planning products.
FRD 5.4.1.4.2	The GSSC shall interface with the ISOC for the exchange of mission planning products.
FRD 5.4.1.4.3	The GSSC shall interface with the GIOC for the exchange of mission planning products.
FRD 5.7.1.2	The GSSC shall receive and archive reports and analyses from the MOC.
FRD 5.7.1.3	The GSSC shall interface with the ISOC for the exchange of data products.
FRD 5.7.1.4	The GSSC shall interface with the GIOC for the exchange of data products.
FRD 5.7.1.6	The GSSC shall receive data products from the MOC, ISOC or GIOC as follows: …
FRD 5.7.1.7	The GSSC shall maintain the integrity of science data received from the IOCs.

Specific Data Product Requirements

Specific Data Product Requirements	
Requirement	**Description**
FRD 5.4.1.4.10	**The GSSC shall receive the TDRSS Contact Schedule Request from the MOC.**
FRD 5.4.1.4.11	**The GSSC shall receive the Confirmed TDRSS Contact Schedule Request from the MOC.**
FRD 5.4.1.4.12	**The GSSC shall receive the Integrated Observatory Timeline from the MOC.**
FRD 5.4.1.4.13	**The GSSC shall receive the orbit data products from the MOC.**
FRD 5.4.1.4.14	**The GSSC shall receive the as-flown timeline from the MOC.**
FRD 5.4.1.5.1	**The GSSC shall receive TOO requests from the science community**
FRD 5.4.1.5.4	**The GSSC shall receive from the MOC information that specifies the status of the TOO order.**
FRD 5.4.1.5.6	**The GSSC shall receive TOO execution notification from the MOC.**
FRD 5.4.1.6.1	**The GSSC shall receive Absolute Time Commands from the IOCs.**
FRD 5.4.1.6.2	**The GSSC shall receive Real Time Commands and File Loads from the IOCs.**
FRD 5.4.1.6.3	**The GSSC shall pass high priority commands, as identified by the IOCs, to the MOC immediately**
FRD 5.4.1.6.6	**The GSSC shall support autonomous data transfers to and from the MOC**
FRD 5.7.1.1	**The GSSC shall receive and archive Level 0 data from the MOC**
FRD 5.7.1.8	**The GSSC shall receive GLAST-produced GCN Notices and Circulars from the GCN.**
SAEDR 4.4.1.2	**Must be able to ingest spacecraft livetime history tables generated by the ISOC for roughly 5 hour periods delivered 5 times per day**
SAEDR 4.4.2.2	**Must be able to ingest the latest complete source catalog updated when necessary by the ISOC**
SAEDR 4.4.3.2	**Must be able to ingest new pulsar ephemerides on update. Must also be able to ingest tables with varying numbers of pulsar ephemerides.**

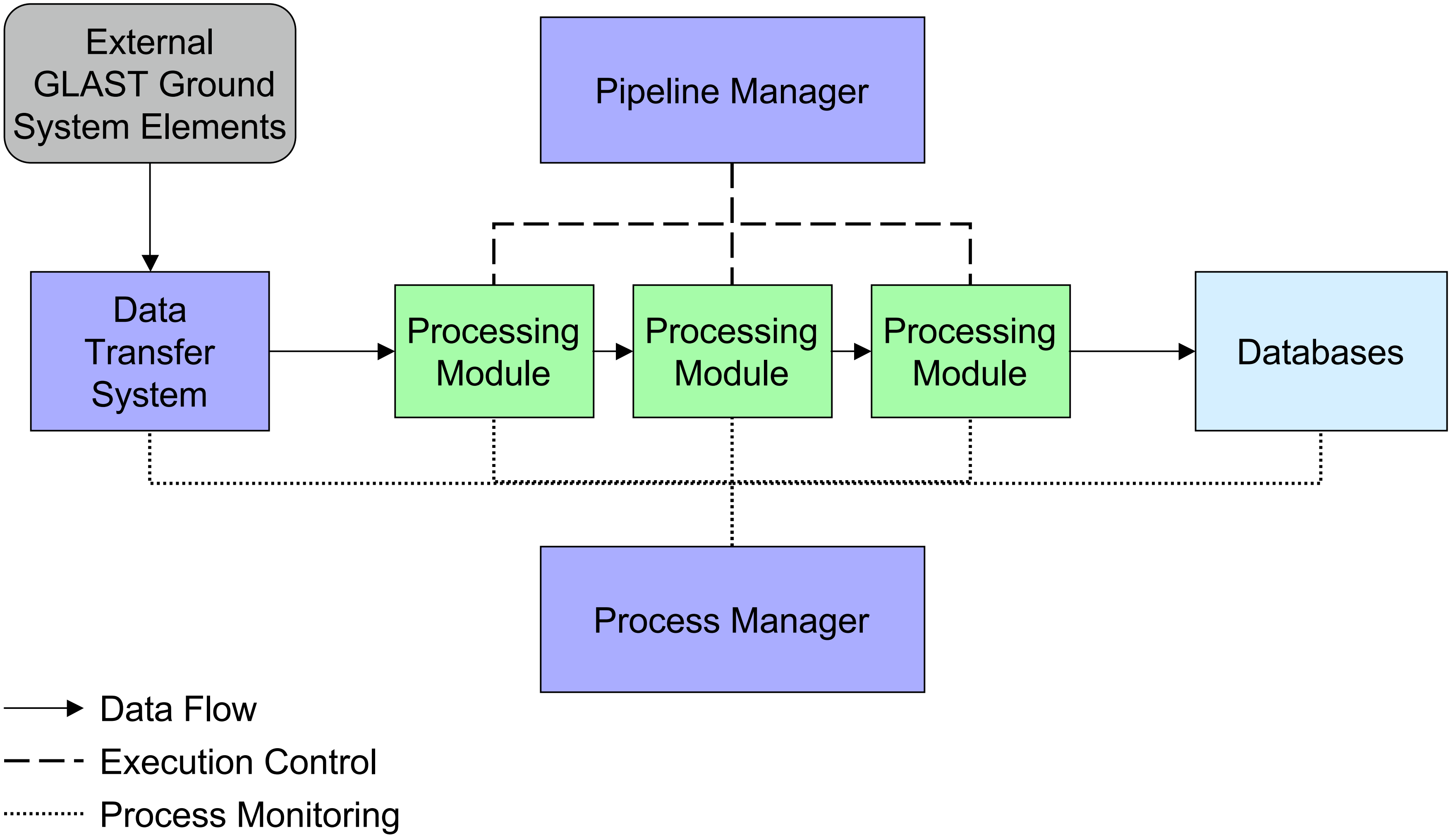
Throughput and Speed Requirements

Throughput and Speed Requirements	
Requirement	Description
FRD 5.7.1.5.1	The GSSC shall be capable of receiving, processing, and archiving Level 0 and Level 1 data resulting from a single downlink of at least 36 hours of observatory data.
FRD 5.7.1.5.2	The GSSC shall be capable of receiving, processing and archiving the Level 0 and Level 1 data generated at an orbit-averaged rate of 1.2 Mbps for LAT, 12 kbps for GBM and 51 kbps for observatory housekeeping data.
LESDR 5.2.3.3	Data must be available for searching in the database 10 minutes after the start of the ingest process for a newly delivered photon summary file assuming the file contains no more than 5 hours worth of data.
LESDR 5.2.3.4.2	A reprocessed 5 hour time photon summary file must be able to be inserted into the database without undue interruption. (It must take less tan 60 minutes to make the new version of the data available for that interval.)
LESDR 5.2.4.3	Data must be available for searching in the database 100 minutes after the start of the ingest process for a newly delivered event summary file assuming the file contains no more than 5 hours worth of data.
LESDR 5.2.4.4.2	Reprocessing an existing 5 hour time interval of data must be done without undue interruption (<10 hours)
SAEDR 4.4.1.5.1	Must be able to ingest a newly delivered 5 hour data table in < 1 minute
SAEDR 4.4.1.5.3	Must be able to input a reprocessed 5 hour data table in < 5 times the time it takes to ingest a brand new table.
SAEDR 4.4.2.5.1	10Mb of LAT point source data must be able to be ingested and read for searching in < 10 min
SAEDR 4.4.2.5.3	Must be able to update tables of refined point source entries at < 5 times the ingest rate
SAEDR 4.4.3.5.1	Must be able to ingest 1 Mb worth of pulsar ephemerides tables in < 1 min
SAEDR 4.4.3.5.3	Must be able to load an updated database in < 5 times the ingest speed

General Ingest System Features

- Provides mechanism by which data is transferred to the GSSC and put into databases accessible by the community
- Automated, hands free system to provide rapid processing of data.
- Consists of four main components
  - Data Transfer System – physically moves the data from the elements of the Ground System to the GSSC
  - Process Manager – Monitors hardware and software components of the Ingest system. Corrects errors detected and notifies operator in case of serious failures or problems
  - Pipeline Manager – Controls the processing of data as it arrives
  - Data Processing Modules – Individual modules activated by the Pipeline Manager to process the data. Includes modules to extract metadata, populate databases, data integrity checks, etc. Customized to each specific data product processed.

Ingest System Diagram



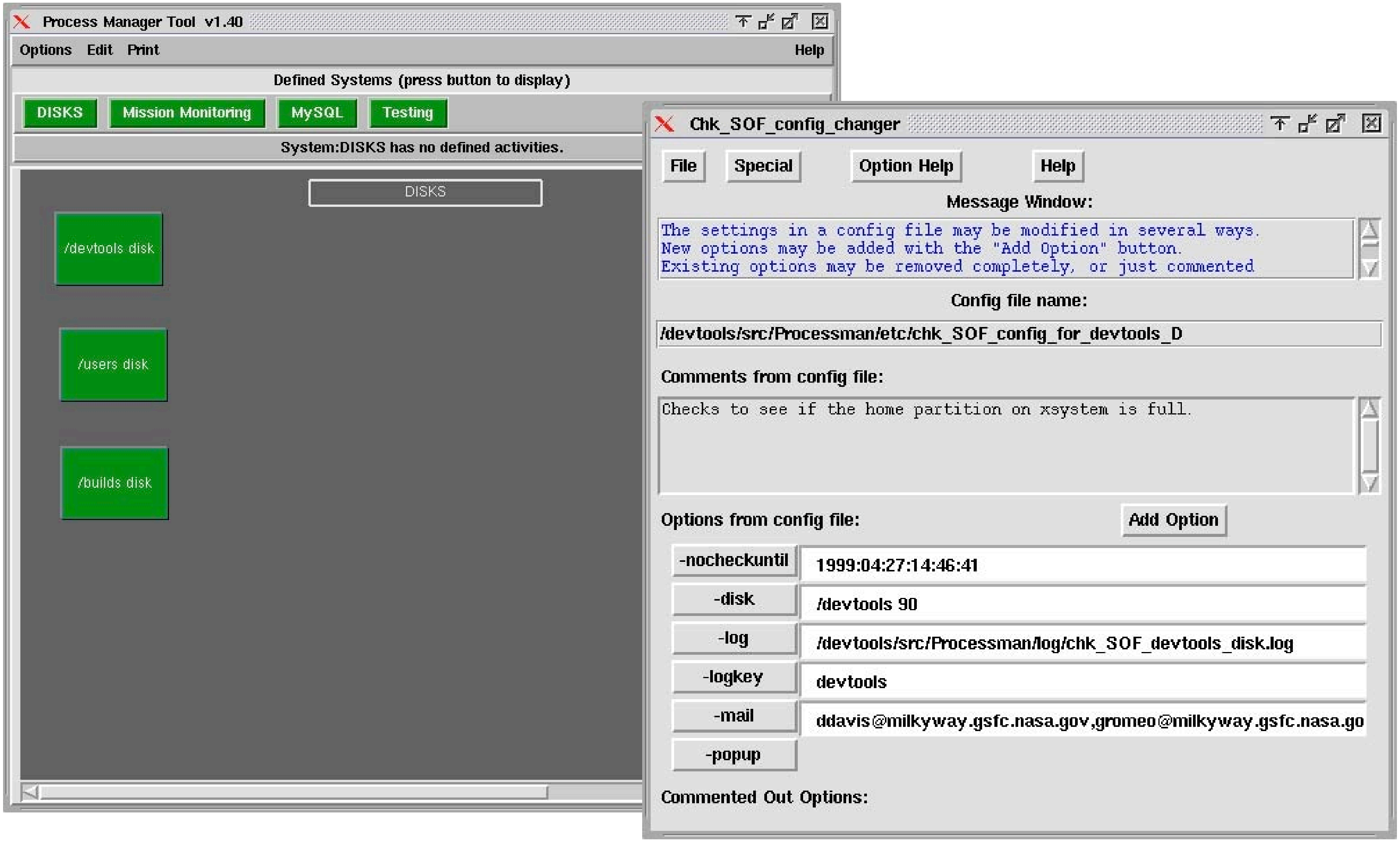
Data Transfer System (DTS)

- Purpose is to provide automated transfer of files between Ground System elements
- The GLAST Ground System will use either DTS or FastCopy as its data transfer system.
  - Provide secure, guaranteed file delivery
  - Allow for post-transfer command execution
- DTS
  - Developed for XMM. Also used by Swift and HEASARC.
  - Maintained by LHEA staff.
  - Uses ssh security protocols
- FastCopy
  - COTS product
  - Private, closed security protocol

Process Manager

- Provides automatic oversight of all processing modules and available resources such as disk space and CPUs.
- Can take corrective action without operator intervention.
- Notifies appropriate parties in the case of serious errors.
- Maintains logs of the state of all processes and resources.
- GSSC will use the Process Manager developed for RXTE with modifications to handle the GSSC's computer system.
  - Back end is a series of scripts that monitor and log system state
  - Front end is a GUI that allows operator to examine and control the state of the system

Process Manager Screen Shots



System:DISKS has no defined activities.

```
The settings in a config file may be modified in several ways.
New options may be added with the "Add Option" button.
Existing options may be removed completely, or just commented
```

/devtools/src/Processman/etc/chk_SOF_config_for_devtools_D

```
Checks to see if the home partition on xsystem is full.
```

-nocheckuntil	1999:04:27:14:46:41
-disk	/devtools 90
-log	/devtools/src/Processman/log/chk_SOF_devtools_disk.log
-logkey	devtools
-mail	ddavis@milkyway.gsfc.nasa.gov,gromeo@milkyway.gsfc.nasa.go
-popup	

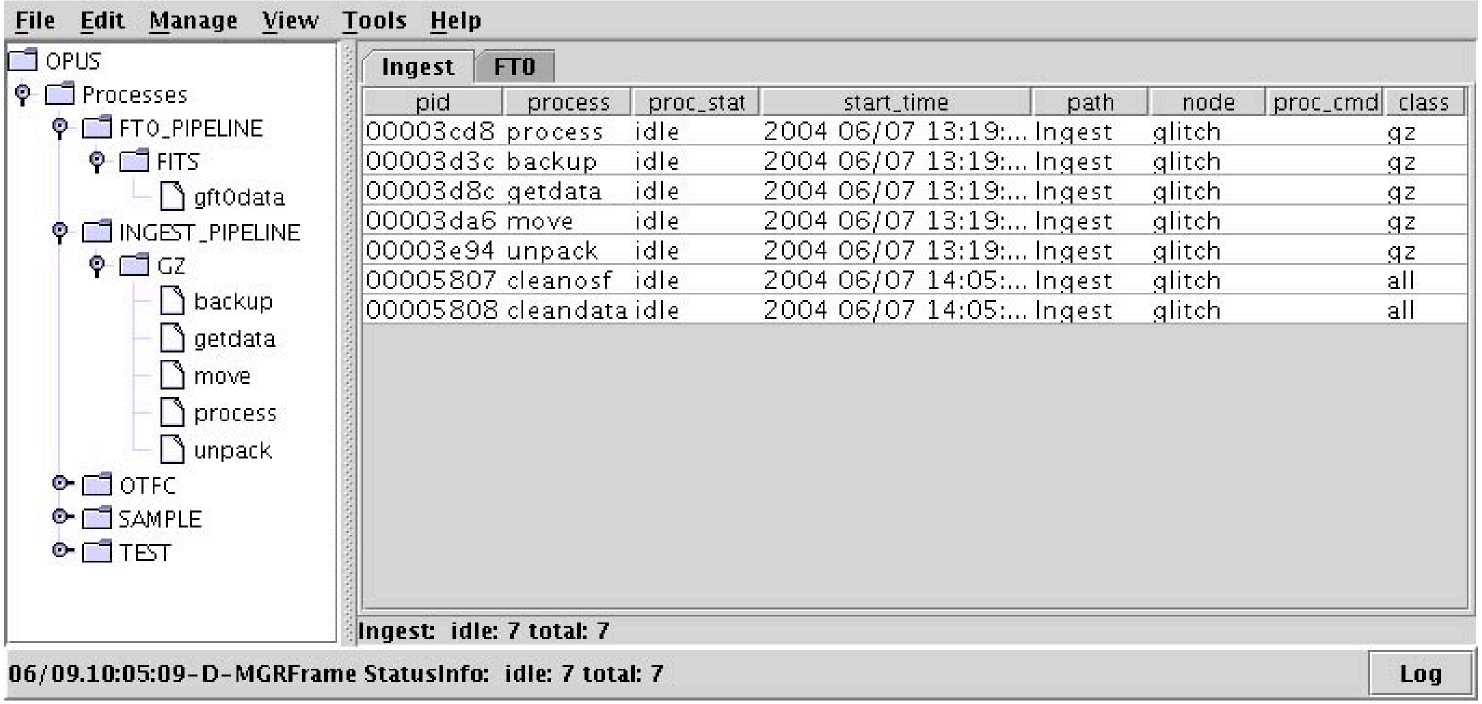
Pipeline Manager (OPUS)

- Controls the processing of data as it arrives and insures that all data is handled properly
- Several options considered and investigated
  - XTE Pipeline manager
  - APS Pipeline manager
  - OPUS
- GSSC has chosen to use OPUS as its Pipeline Manager
  - All choices required a large amount of code development
  - OPUS was easier to integrate and provided pre-built monitoring tools.
  - OPUS currently being used by the ISOC for its Level 1 processing pipeline

OPUS

- Developed at STScI for processing of HST data.
- Lightweight & Flexible
  - OPUS only provides pipeline management backbone
  - Actual processing steps can be arbitrarily simple or complex
- Scalable
  - OPUS is designed to manage multiple instances of multiple processing steps on multiple computers.
  - Additional processing resources can be added “on the fly”.
- Pre-built monitoring GUI’s
  - Process Manager (PMG) – View of processes associated with each branch of the pipeline including number of instances, location of processes and data currently being processed
  - Observation Manager (OMG) – Data oriented view showing which stages of the pipeline have been complete for each data set.
  - Written in Java and can be run from anywhere.

OPUS PMG



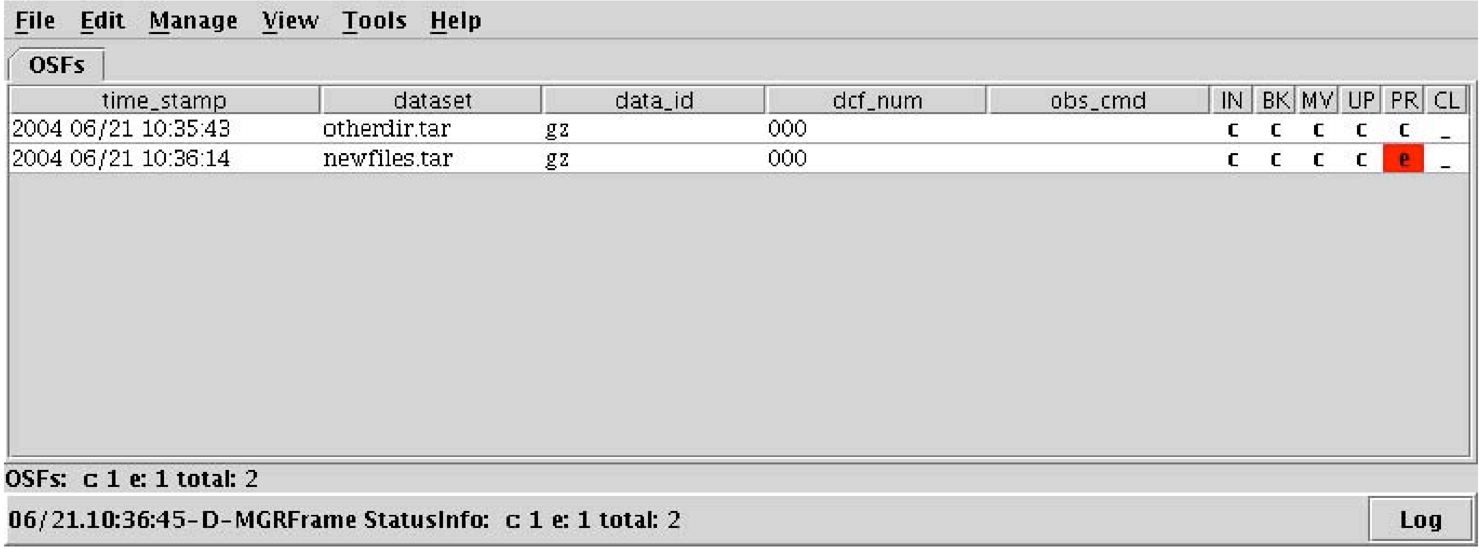
pid	process	proc_stat	start_time	path	node	proc_cmd	class
00003cd8	process	idle	2004 06/07 13:19:...	Ingest	glitch		gz
00003d3c	backup	idle	2004 06/07 13:19:...	Ingest	glitch		gz
00003d8c	getdata	idle	2004 06/07 13:19:...	Ingest	glitch		gz
00003da6	move	idle	2004 06/07 13:19:...	Ingest	glitch		gz
00003e94	unpack	idle	2004 06/07 13:19:...	Ingest	glitch		gz
00005807	cleanosf	idle	2004 06/07 14:05:...	Ingest	glitch		all
00005808	cleandata	idle	2004 06/07 14:05:...	Ingest	glitch		all

Ingest: idle: 7 total: 7

06/09.10:05:09-D-MGRFrame StatusInfo: idle: 7 total: 7

- Process oriented view
- Shows process name, host computer, current data set/status
- One view per pipeline branch

OPUS OMG



time_stamp	dataset	data_id	dcf_num	obs_cmd	IN	BK	MV	UP	PR	CL
2004 06/21 10:35:43	otherdir.tar	gz	000		c	c	c	c	c	_
2004 06/21 10:36:14	newfiles.tar	gz	000		c	c	c	c	e	_

OSFs: c 1 e: 1 total: 2

06/21.10:36:45-D-MGRFrame StatusInfo: c 1 e: 1 total: 2

- Data oriented view – one entry per data item
- Shows status of each step of the pipeline branch for each data set
- Errors flagged in red
- One view for each pipeline branch

Data Tracking

- Each step of processing is logged in a MySQL database

- Database tables log start and completion time for each stage of the processing pipeline for each file that moves through the system.

- Important metadata are also tracked, including data contents, revision numbers, etc.

Error Handling

- When there is an error, OPUS flags the data set that had the error and the processing step in which the error occurred

- Additional portions of the pipeline system monitor all data for error conditions.

- Appropriate actions taken based on error type. Some examples are:
  - “Known” processing error
  - Unexpected processing error
  - Crashed process

- All errors are logged

Specific Pipeline Branches

- Some processing steps are the same for every data product received.

- Each data product also has specific, individual processing branch that needs to be followed
  - Many of the steps on these branches are dependent on the data format.
  - These will be completed as the detailed definition of the data format they work on are developed.

- These branches are represented by “paths” in the OPUS framework

General Ingest Branch

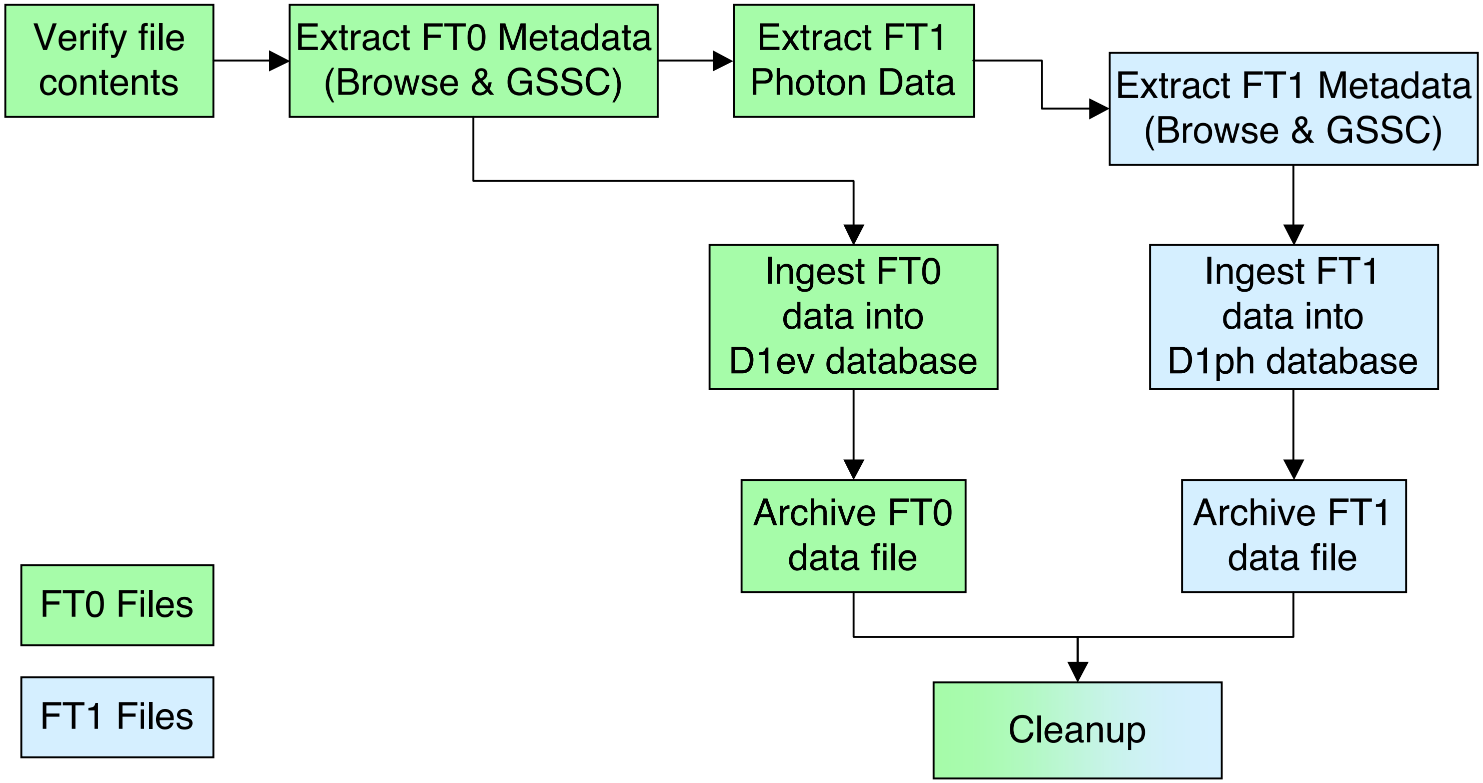
- Performed on all the data received by the GSSC
- Consists of 5 main steps
  - Receive data – Gets the data from the data transfer system and starts the pipeline going
  - Backup data – Makes a copy of the data to the GSSC data backup system
  - Unpack data – Unpacks the data from the compressed form used for the data transfer
  - Identify data products – Identify each individual data product delivered in the transfer and move it into the appropriate data ingest pipeline branch. Identification done by parsing the header files accompanying each data product.
  - Clean-up – Clean up temporary files and directories used by this portion of the ingest process

Level 0 Data Pipeline

- Pipeline to handle the Level 0 data received from the MOC
- Simplest of all pipelines since the data is simply archived and not part of a public database
- Only 3 steps
  - Verify data – Verify data integrity
  - Extract Metadata – Extract the metadata from the data header to monitor the data that has been received
  - Archive Data – Add the new data to the permanent Level 0 data archive

LAT Event Data Pipeline

- Handles ingest of level 1 LAT Event Data
- One of the more complex processing pipelines
- Consists of 9 main steps


Verify file contents
Extract FT0 Metadata (Browse & GSSC)
Extract FT1 Photon Data
Extract FT1 Metadata (Browse & GSSC)
Ingest FT0 data into D1ev database
Ingest FT1 data into D1ph database
Archive FT0 data file
Archive FT1 data file
FT0 Files
FT1 Files
Cleanup

Data Ingest Completion Milestones

- OPUS is already installed and running
- Initial Ingest branch designed and prototype has been implemented
- Data specific module completion tied to GSSC Software Releases
  - Release 1 (11/15/04) – Data Transfer System, OPUS, Initial Ingest branch and Level 0 specific Pipeline
  - Release 2 (02/01/05) – Operations Data I
  - Release 3 (05/01/05) – Operations Data II
  - Release 4 (08/01/05) – GBM Science Data, LAT Science Data, Operations Data III
  - Release 5 (01/31/06) – Anomaly Reports pipeline and Process Manager
  - Release 6 (04/03/06) – All remaining data
  - Release 7 (01/15/07) – No new modules

Backup Level 1 Pipelines

- Primary resource driver is the LAT level 1 processing pipeline
  - Currently requires 1080 CPU-hours on a ~2 GHz CPU to process 1 day's worth of LAT data
  - One day's data generates ~52 GB of final Level 1 data and ~700 GB of intermediate data
- Pursuing partial buy-in to an existing or planned cluster here at GSFC
  - Thunderhead – 512 2.4GHz Xeon processors
  - Bliss – 140 2.2GHz Xeon processors
  - Bliss II – to be purchased by Dec '05.
- Should rarely if ever be used for pipeline processing – free time to be available to LHEA and the gamma ray community for scientific studies.
- To be setup by GSSC Software Release 5 (01/31/06) in preparation for GRT 6 (03/15/06).

Summary

- GSSC hosts 3 different pipelines
- Ingest pipeline handles all data coming into the GSSC
- OPUS will serve as the Pipeline Manager for the GSSC
- OPUS and initial ingest pipeline prototype installed and running
- All data will be tracked in an Ingest database
- Plans for procuring necessary hardware for backup instrument data processing pipelines well underway.